Light-harvesting chlorophyll a/b-binding protein-coding genes in jatropha and the comparison with castor, cassava and arabidopsis

Yongguo Zhao [1, 2] , **Hua Kong** [2] , **Yunling Guo** [2] , **Zhi Zou** [Corresp. 2]

[1] Guangdong University of Petrochemical Technology, Maoming, China

[2] Key Laboratory of Biology and Genetic Resources of Tropical Crops, Ministry of Agriculture and Rural Affairs, Hainan Key Laboratory for Biosafety Monitoring and Molecular Breeding in Off-Season Reproduction Regions, Institute of Tropical Biosciences and Biotechnology, Chinese Academy of Tropical Agricultural Science, Haikou, China

Corresponding Author: Zhi Zou
Email address: zouzhi@itbb.org.cn

The Lhc (light-harvesting chlorophyll a/b-binding protein) superfamily represents a class of antennae proteins that play indispensable roles in capture of solar energy as well as photoprotection under stress conditions. Despite their importance, little information has been available beyond model plants. In this study, we presents a first genome-wide analysis of *Lhc* superfamily genes in jatropha (*Jatropha curcas* L., Euphorbiaceae), an oil-bearing plant for biodiesel purpose. A total of 27 members were identified from the jatropha genome, which were shown to distribute over nine out of the 11 chromosomes. The superfamily number is comparable to 28 present in castor (*Ricinus communis*, Euphorbiaceae), but relatively less than 35 in cassava (*Manihot esculenta*, Euphorbiaceae) and 34 in arabidopsis (*Arabidopsis thaliana*) that experienced one or two recent whole-genome duplications (WGDs), respectively. In contrast to a high number of paralogs present in cassava and arabidopsis, few duplicates were found in jatropha as observed in castor, corresponding to no recent WGD occurred in these two species. Nevertheless, 26 orthologous groups representing four defined families were found in jatropha, and nearly one-to-one orthologous relationship was observed between jatropha and castor. By contrast, a novel group named SEP6 was shown to have been lost in arabidopsis. Global transcriptome profiling revealed a predominant expression pattern of most *JcLhc* superfamily genes in green tissues, reflecting their key roles in photosynthesis. Moreover, their expression profiles upon hormones, drought, and salt stresses were also investigated. These findings not only improve our knowledge on species-specific evolution of the *Lhc* supergene family, but also provide valuable information for further studies in jatropha.

1 **TITLE:**

2 **Light-harvesting chlorophyll a/b-binding protein-coding genes in jatropha**

3 **and the comparison with castor, cassava and arabidopsis**

6 **AUTHORS:**

7 Yongguo Zhao[1,2], Hua Kong[2], Yunling Guo[2], Zhi Zou[2,*]

10 **E-MAIL ADDRESSES:**

11 zhaoyongguo@gdupt.edu.cn (Y.G Zhao), konghua@itbb.org.cn (H. Kong), guoyunling@itbb.org.cn (Y.L.

12 Guo), zouzhi2008@126.com or zouzhi@itbb.org.cn (Z. Zou)

15 **AFFILIATIONS:**

16 [1] Guangdong University of Petrochemical Technology, Maoming, Guangdong, China

17 [2] Hainan Key Laboratory for Biosafety Monitoring and Molecular Breeding in Off-Season Reproduction

18 Regions, Key Laboratory of Biology and Genetic Resources of Tropical Crops, Ministry of Agriculture and

19 Rural Affairs, Institute of Tropical Biosciences and Biotechnology, Chinese Academy of Tropical Agricultural

20 Sciences, Haikou, Hainan, China

23 [*]To whom correspondence should be addressed. E-mail: zouzhi2008@126.com or zouzhi@itbb.org.cn.

24 ORCID ID: 0000-0003-2881-1932

27 **RUNNING TITLE:**

28 *Lhc* superfamily genes in jatropha

29 **ABSTRACT**

30 The Lhc (light-harvesting chlorophyll a/b-binding protein) superfamily represents a class of antennae proteins

31 that play indispensable roles in capture of solar energy as well as photoprotection under stress conditions.

32 Despite their importance, little information has been available beyond model plants. In this study, we presents

33 a first genome-wide analysis of *Lhc* superfamily genes in jatropha (*Jatropha curcas* L., Euphorbiaceae), an oil-

34 bearing plant for biodiesel purpose. A total of 27 members were identified from the jatropha genome, which

35 were shown to distribute over nine out of the 11 chromosomes. The superfamily number is comparable to 28

36 present in castor (*Ricinus communis*, Euphorbiaceae), but relatively less than 35 in cassava (*Manihot esculenta*,

37 Euphorbiaceae) and 34 in arabidopsis (*Arabidopsis thaliana*) that experienced one or two recent whole-

38 genome duplications (WGDs), respectively. In contrast to a high number of paralogs present in cassava and

39 arabidopsis, few duplicates were found in jatropha as observed in castor, corresponding to no recent WGD

40 occurred in these two species. Nevertheless, 26 orthologous groups representing four defined families were

41 found in jatropha, and nearly one-to-one orthologous relationship was observed between jatropha and castor.

42 By contrast, a novel group named SEP6 was shown to have been lost in arabidopsis. Global transcriptome

43 profiling revealed a predominant expression pattern of most *JcLhc* superfamily genes in green tissues,

44 reflecting their key roles in photosynthesis. Moreover, their expression profiles upon hormones, drought, and

45 salt stresses were also investigated. These findings not only improve our knowledge on species-specific

46 evolution of the *Lhc* supergene family, but also provide valuable information for further studies in jatropha.

47 **Keywords:** *Jatropha curcas*, *Ricinus communis*, *Manihot esculenta*, *Arabidopsis thaliana*, Light-harvesting

48 chlorophyll a/b-binding protein, *Lhc* superfamily, Evolution, Synteny analysis, Whole-genome duplication,

49 Expression profile

50 **INTRODUCTION**

51 Light-harvesting chlorophyll a/b-binding protein (Lhc) superfamily, defined by the presence of a conserved

52 chlorophyll-binding (CB) domain in the transmembrane alpha-helix, is composed of four distinct nuclear-

53 encoded antennae protein families in green plants, i.e. Lhc, Lil (light-harvesting-like), PsbS (photosystem II

54   subunit S), and FCII (ferrochelatase II) (*Klimmek et al., 2006; Zou & Yang, 2019a*). In contrast to an orphan

55   group present in both PsbS and FCII families, the Lhc family, initially known as CAB (chlorophyll a/b-binding

56   protein), contains two evolutionary groups named Lhca and Lhcb that are associated with photosystem I or II

57   (PSI/II), respectively (*Jansson, 1999; Klimmek et al., 2006*). The Lil family includes four diverse subfamilies,

58   i.e. OHP (one-helix protein), SEP (stress-enhanced protein), Lil1 or ELIP (early light-induced protein), and

59   Lil8 or Psb33 (photosystem II protein 33) (*Engelken et al., 2010; Zou, 2018*). Among them, OHP and SEP can

60   be further divided into several groups: the OHP subfamily includes two groups named OHP1/Lil2 and

61   OHP2/Lil6, whereas the SEP subfamily contains six groups, i.e. SEP1/Lil4, SEP2/Lil5, SEP3/Lil3, SEP4,

62   SEP5, and SEP6 (*Engelken et al., 2010; Zou & Yang, 2019a*). Investigation of their origin suggested that OHP

63   is more primitive, which is more likely to result from the plastid-encoded HLIP (high light-induced protein)

64   via gene transfer after the primary endosymbiosis (*Koziol et al., 2007; Engelken et al., 2010*). In addition to

65   light harvesting and transport, growing evidence shows that *Lhc* superfamily members are also involved in

66   regulation and distribution of excitation energy between PSI and PSII, maintenance of thylakoid membrane

67   structure, photoprotection as well as response to various stresses (*Pan et al., 2011; Fan et al., 2015; Fristedt et*

68   *al., 2015; Hey et al., 2017; Myouga et al., 2018*).

69   *Jatropha curcas* L. ($2n$ = 22), commonly known as jatropha, physic nut, barbados nut, or purging nut, is a

70   perennial large shrub or small tree (*Mazumdar et al., 2018; Zou et al., 2016, 2018*). Jatropha belongs to the

71   Euphorbiaceae family, which also includes castor (also known as castor bean, *Ricinus communis* L.) and

72   cassava (*Manihot esculenta* Crantz) and is characterized with high photosynthesis and high biomass (*Zou et al.,*

73   *2018; Zou & Yang, 2019a, 2019b*). As a potential non-edible energy crop, jatropha produces high level of

74   fossil fuel-like oil in its seeds, which can be easily processed into biodiesel (*Fairless, 2007; Berchmans &*

75   *Hirata, 2008; Kumar & Sharma, 2008; Maghuly & Laimer, 2013*). Additionally, this species also has several

76   unique characteristics like easy propagation, rapid growth, and adaptation to semiarid and barren soil

77   environments (*Montes & Melchinger, 2016*). Although originated from Mesoamerica, jatropha can now be

78   widely found in many tropical and subtropical countries of Africa and Asia (*Wu et al., 2015; Li et al., 2017*).

79   Nevertheless, its commercial cultivation has failed mainly due to low productivity (*Montes & Melchinger,*

80   *2016; Mazumdar et al., 2018*). Thereby, uncovering the molecular mechanism underlying and characterization

81 of genes involved in yield formation are prerequisites. In this study, we would like to present a first genome-

82 wide analysis of the *Lhc* supergene family in jatropha, including gene structures, chromosome (Chr) locations,

83 evolutionary relationships, sequence characteristics, global expression profiles as well as comprehensive

84 comparison with arabidopsis, cassava, and castor. These results will not only improve our knowledge on

85 species-specific evolution of the *Lhc* supergene family, but also provide valuable information for further

86 functional analysis in jatropha.

87 **MATERIALS & METHODS**

88 **Identification of *Lhc* superfamily genes**

89 As shown in **Supplementary Table S1**, 34 *AtLhc* superfamily genes were retrieved from TAIR

90 (https://www.arabidopsis.org/, Araport11) according to previous literatures. To facilitate evolutionary analysis,

91 28 and 35 superfamily members present in castor and cassava (see **Supplementary Table S1**), two

92 Euphorbiaceous plants, were also obtained from Phytozome (Version 12,

93 https://phytozome.jgi.doe.gov/pz/portal.html). Homologs present in the jatropha genome (*Wu et al., 2015*)

94 were identified via the tBLASTn (*Altschul et al., 1997*; *E*-value, 1e–5) search by using above protein

95 sequences as queries. Gene models of candidates were revised via aligning mRNA to loci-encoding scaffolds.

96 Presence of the conserved CB domain (PF00504) was checked using MOTIF Search

97 (https://www.genome.jp/tools/motif/), and exon-intron structures were displayed using Gene Structure Display

98 Server (GSDS 2.0, https://gsds.cbi.pku.edu.cn/). Putative transmembrane helix (TMH) was predicted using

99 CCTOP (http://cctop.enzim.ttk.mta.hu/) as well as sequence alignment. Chloroplast transit peptide (TP) of

100 deduced proteins and biochemical parameters of mature peptides were determined using ChloroP (Version 1.1,

101 https://www.cbs.dtu.dk/services/ChloroP/) and ProtParam (https://web.expasy.org/protparam/), respectively.

102 **Chromosome location and synteny analysis**

103 Gene distribution on chromosomes was analyzed using MAPchart 2.3 (*Voorrips, 2002*). For synteny analysis,

104 the all-to-all BLASTP method was used to identify duplicate pairs, and MicroSyn (*Cai et al., 2011*) was used

105 to detect microsynteny. Orthologs across different species were inferred from the best-reciprocal-hit (BRH)-

106 based BLAST analysis as well as synteny analysis for jatropha and castor.

107 **Sequence alignment, phylogenetic, and conserved motif analyses**

108   Multiple sequence alignment was carried out using MUSCLE (*Edgar, 2004*). Phylogenetic tree construction

109   was performed using MEGA7 (*Tamura et al., 2016*) with the maximum likelihood method (bootstrap: 1,000).

110   Conserved motifs were identified using MEME (https://meme-suite.org/tools/meme): any number of

111   repetitions; maximum number of motifs, 25; minimum sites, 2; and, the optimum width of each motif, between

112   6 and 100 residues.

113   **Gene expression analysis**

114   Transcriptome datasets used for expression profiling are shown in **Supplementary Table S2**. Except for

115   tissue-specific transcriptomes, other samples were performed for at least two biological replicates. Quality

116   control   of   raw   reads   was   carried   out   using   fastQC   (https://www.

117   bioinformatics.babraham.ac.uk/projects/fastqc/). Read mapping were performed using Bowtie 2 (*Langmead &*

118   *Salzberg, 2012*), and the relative transcript level of each gene was presented as FPKM (fragments per kilobase

119   of exon per million fragments mapped, for pair-ended samples) or RPKM (Reads per kilobase per million

120   mapped reads, for single-ended samples) (*Mortazavi et al., 2008*). RSEM (v1.2.27) (*Li & Dewey, 2011*) with

121   parameters "log2Ratio $\geq$ 1" and "FDR < 0.001" were used to determine differentially expressed genes.

122   **RESULTS**

123   **Identification and chromosome locations of 27 *Lhc* superfamily genes in jatropha**

124   The BLAST search resulted in 27 *JcLhc* superfamily genes from the jatropha genome (*Wu et al., 2015*), which

125   represent four previously defined families (i.e. *Lhc*, *Lil*, *PsbS*, and *FCII*) or eight subfamilies (i.e. *Lhca*, *Lhcb*,

126   *PsbS*, *OHP*, *SEP*, *ELIP*, *Psb33*, and *FCII*). Each subfamily contains one to nine members that were named

127   after their orthologs in castor (see below), i.e. *JcLhca1–6*, *JcLhcb1.1–1.2*, *JcLhcb2–8*, *JcPsbS*, *JcELIP*,

128   *JcOHP1–2*, *JcSEP1–6*, *JcPsb33*, and *JcFCII*, respectively. These genes were shown to distribute over 25

129   scaffolds. Although most scaffolds contain a single member, two of them harbor two, i.e. scaffold160 (i.e.

130   *JcLhcb1.1* and *JcLhcb1.2*) and scaffold211 (i.e. *JcLhca6* and *JcSEP5*) (**Table 1**). With the help of 1,208

131   available genetic markers, these genes were further anchored to nine chromosomes, and the gene number of

132   each chromosome varies from one to five (**Fig. 1**).

133   The expression of all *JcLhc* genes was supported by available Sanger sequencing-derived expressed

134   sequence tags (ESTs) and/or RNA sequencing (RNA-seq), where *JcLhcb1.1* harbors the maximum of 28 EST

135   hits. The intron number of these genes varies from zero to nine: approximately 11.11% of genes are intronless,

136   and 29.63%, 22.22%, 11.11%, 11.11%, 11.11% or 3.70% contain two, one, three, four, five and nine introns,

137   respectively (**Table 1** and **Fig. 2**). Similar exon-intron structure was also observed in castor, cassava, and

138   arabidopsis (see **Supplementary Table S1**), implying a conserved evolution between these species. The

139   average length of coding sequences (CDS) is about 760 bp, varying from 357 bp of *JcOHP1* to 1,500 bp of

140   *JcFCII*. Compared with CDS, the intron length is relatively more variable, ranging from 79 bp of *JcLhcb6* to

141   8,100 bp of *JcFCII* and with the average length of 1,259 bp (**Table 1** and **Fig. 2**). The CDS of *JcLhcb1.1* and

142   *JcLhcb1.2*, which are reversely clustered on scaffold160, was shown to exhibit 96.7% identity. Thereby, they

143   are more likely to result from tandem duplication (*Zou et al., 2019; Zou & Yang, 2019a, 2019c*).

144   ***Synteny analysis and determination of orthologous groups***

145   Orthologs of *JcLhc* superfamily genes in castor, cassava, and arabidopsis were further identified by using the

146   BRH method, resulting in 26 orthologous groups (OGs) when the definition was confined to at least one

147   member present in more than two species examined (**Table 1**). The result is highly consistent with

148   phylogenetic analysis (see below) as well as synteny analysis performed between jatropha and castor, where

149   one-to-one orthologous relationship was observed with exception of the Lhcb1 group with two-to-three (**Table**

150   **1**). Interestingly, *RcLhcb1.2*, which may originate by dispersed duplication, is located on scaffold30005

151   together with *RcLhca3*. However, no collinear gene was found for *RcLhcb1.2* in jatropha (see **Supplementary**

152   **Fig. S1**). By contrast, orthologous relationships between jatropha and cassava/arabidopsis are relatively

153   complex, which include one-to-one, one-to-two, one-to-three, and two-to-five, corresponding to one or more

154   recent WGDs occurred in these two species. It is worth noting that, SEP6, a recently identified group that is

155   present in jatropha, castor, and cassava, is absent from arabidopsis (**Table 1**), implying species/lineage-specific

156   gene loss. Additionally, species-specific gene expansion was also observed: duplicates identified in jatropha

157   (one) and castor (two) were shown to result from tandem or dispersed duplication, respectively; nine duplicates

158   identified in cassava were derived from tandem duplication (three) and whole-genome duplication (WGD)

159   (six); in arabidopsis, four or five duplicates were derived from tandem duplication and WGD, respectively

160   (**Table 1** and **Supplementary Table S1**).

161   Although not exactly the same within a family, the exon-intron structure is highly conserved within a certain

162   OG: Lhcb1 and SEP4 are intronless; one-intron-containing groups include Lhcb2/-4/-6/-8, OHP2, and SEP2,

163   whereas two-intron groups include Lhca3/-4, Lhcb3, ELIP, OHP1, SEP3/-6, and Psb33; Lhca1 and SEP1

164   feature three introns, whereas Lhca2/-6, and SEP5 feature four introns; three groups (i.e. Lhcb5/-7 and Lhca5)

165   contain five introns, whereas only one group (i.e. FCII) harbors nine introns (**Fig. 2**, **Table 1**, and

166   **Supplementary Table S1**).

167   **Phylogenetic analysis, sequence features, and conserved motifs**

168   As shown in **Table 1**, the deduced JcLhc superfamily proteins consist of 118–499 amino acids (AA) with one

169   to four TMHs (**Supplementary Fig. S2**), and the predicted length of transit peptide ranges from 31 to 100

170   residues (**Table 1**). Several physical and chemical parameters of mature peptides were further calculated: the

171   molecular weight (MW) and isoelectric point (pI) values of mature proteins in jatropha range from 7.55 to

172   46.76 kilodalton (kDa) or from 4.56 to 9.99, respectively; about 81.48% of JcLhc superfamily proteins harbor

173   a pI value of less than 7, which is relatively less than 85.71% in castor, 91.43% in cassava, or 91.18% in

174   arabidopsis; and, about 59.26% of JcLhc superfamily proteins harbor a grand average of hydropathicity

175   (GRAVY) value of less than 0, which is relatively less than 64.29% in castor, 62.86% in cassava, or 73.53% in

176   arabidopsis (**Table 1** and **Supplementary Table S1**).

177   Except for JcPsb33 that contains a CB-like domain (see **Supplementary Fig. S3**), all other JcLhc

178   superfamily proteins include the core CB domain of approximately 20 AA (see **Supplementary Table S1**).

179   Nevertheless, their overall sequence similarity was shown to be considerably low, even within the conserved

180   Lhc family (ranging from 27.2% to 98.5%, see **Supplementary Table S3**). To keep the analysis reliable, an

181   independent phylogenetic tree was constructed for each subfamily by using full-length proteins from jatropha,

182   castor, cassava, and arabidopsis. As shown in **Fig. 3a**, subfamilies Lhca, Lhcb, OHP, and SEP are clearly

183   clustered into six, eight, two or six phylogenetic groups respectively, corresponding to 22 OGs as described

184   above, i.e. Lhca1–6, Lhcb1–8, OHP1, OHP2, and SEP1–6. Among them, Lhcb8 and SEP6 exhibit a closer

185   relationship with Lhcb4 and SEP3, with a similarity of 79.5% or 55.5%, respectively (**Supplementary Table**

186   **S3**), where JcLhcb8 harbors a relative shorter C-terminal in relation to JcLhcb4 (**Supplementary Fig. S2**).

187   To reveal possible divergence of members within a certain OG and between different OGs/(sub)families,

188   conserved motifs were analyzed using MEME. As shown in **Fig. 3b** and **Table 2**, motifs are highly variable

189   between subfamilies or even between different evolutionary groups, and considerably more motifs were

190   identified in the Lhc family as compared with PsbS, FCII, and Lil families. Among 25 motifs identified, Motif

191   1, which is characterized as the CB domain, is widely distributed, including Psb33s. Motif 19, which is

192   characterized as chloroplast transit peptide, is also widely found. Motifs 25 and 11 are characterized as part of

193   the Rieske-like domain (PF13806), where Motif 11 is Psb33-specific and Motif 25 is also present in Lhca1s.

194   The Ferrochelatase domain (PF00762), which is FCII-specific, was shown to include Motifs 21, 12, 10, and 24.

195   Among them, Motif 10 is also present in SEP3s and SEP6s. Motif 7, which includes the WYGPDR/WYGEER

196   domain, is widely found in Lhcb1, -2, -3, -5, and -7 groups. WYGPDR has been proven to be essential for

197   trimerization (*Hobe et al., 1995; Rogl & Kühlbrandt, 1999*), however, experimental evidence is still needed for

198   WYGPD and WYGEER varieties. By contrast, little information is available for other motifs, including

199   several group-specific motifs such as Motifs 8, 9, 18, 20, and 22.

200   Although these motifs are usually conserved within a certain OG, species-specific gain or loss was also

201   observed. For example, the conserved Motifs 15, 23, and 10 are absent from AtLhca5, AtLhca3 or AtFCII,

202   respectively, whereas the widely present Motif 8 in SEP3s and SEP6s is replaced by Motif 3 in MeSEP6 (**Fig.**

203   **3**).

204   **Expression profiles of *JcLhc* superfamily genes**

205   Global gene expression profiles were investigated in various tissues, i.e. root (from 15-day-old seedlings),

206   leafage (from 4-year-old plants, half expanded), leaf (mature leaf, fully expanded), IND (undifferentiated

207   inflorescence of 0.5 cm diameter), PID1 (female flower with carpel primordia beginning to differentiate), PID2

208   (female flower with three distinct carpels formed), STD1 (male flower with stamen primordia beginning to

209   differentiate), STD2 (male flower with ten complete stamens formed), and developing seed (19−28 days after

210   pollination). Despite the expression of all identified genes, their transcript levels are highly variable over

211   different tissues. As shown in **Fig. 4**, the majority of *JcLhc* superfamily genes are predominantly expressed in

212   leaf, and the total transcript level of the whole superfamily in STD1, PID2, STD2, PID1, leafage, root, and

213   seed accounts for 42.82%, 24.26%, 23.71%, 21.41%, 21.16%, 18.01%, 5.83%, or 1.75% of that in leaf,

214 respectively. According to tissue-specific expression patterns, *JcLhc* superfamily genes can be divided into

215 five main clusters: Cluster I is mostly expressed in leafage, including *JcPsbS*, *JcFCII*, and four *Lils* (i.e.

216 *JcOHP2*, *JcSEP1*, *JcSEP2*, and *JcELIP*); Clusters II and III that contain the vast majority of *Lhc* family

217 members and include approximately 63.00% of the whole superfamily are predominantly expressed in mature

218 leaf, where Cluster III is also highly abundant in leafage (i.e. *JcLhca6*, *JcLhcb7*, *JcLhcb8*, *JcOHP1*, *JcSEP3*,

219 *JcSEP4*, *JcSEP5*, and *JcPsb33*); Cluster IV that only includes *JcSEP6* is preferentially expressed in root,

220 whereas Cluster V is typically expressed in STD1 (i.e. *JcLhcb1.2*, *JcLhcb2*, and *JcLhcb6*) (**Fig. 4**). It is worth

221 noting that, *JcLhcb1.1* represents the most expressed gene in most tissues examined, implying its key roles.

222 Given drought and salt are two of the most important abiotic stresses affecting plant growth and

223 development, photosynthesis, and crop yield, we thereby investigated the response patterns of *JcLhc*

224 superfamily genes post drought or salt treatment in leaves and roots of eight-week-old seedlings. After

225 withholding irrigation for 1, 4 or 7 d, the total superfamily transcripts in roots were not significantly changed,

226 by contrast, initial increase followed by significant decrease were observed in leaves. The result is consistent

227 with the fact that the net photosynthesis rate (Pn) and stomatal conductance had decreased to 80% or 20% of

228 those in the control after the start of the stress treatment for 2 and 7 d, respectively (*Zhang et al., 2015*). For 1

229 d, seven or three genes were significantly regulated in leaves and roots, respectively. Among them, *JcLhcb1.1*,

230 *JcLhcb1.2*, *JcLhcb3*, *JcELIP*, *JcSEP2*, and *JcSEP5* were upregulated in leaves, whereas *JcOHP2* was

231 downregulated; *JcLhcb3* and *JcFCII* were downregulated in roots, whereas *JcPsbS* was upregulated. For 4 d,

232 ten or seven genes were significantly regulated in leaves and roots, respectively. Among them, *JcLhca4*,

233 *JcLhcb2*, *JcLhcb4*, and *JcELIP* were upregulated in leaves, whereas *JcLhcb1.1*, *JcLhcb1.2*, *JcLhcb3*, *JcLhcb5*,

234 *JcLhcb7*, and *JcLhcb8* were downregulated; *JcLhca3*, *JcLhca4*, *JcLhcb3*, *JcLhcb8*, and *JcPsbS* were

235 upregulated in roots, whereas *JcLhca1* and *JcFCII* were downregulated. For 7 d, 19 or seven genes were

236 significantly regulated in leaves and roots, respectively. Among them, *JcLhca4*, *JcLhcb2*, *JcPsbS*, *JcELIP*,

237 *JcOHP1*, *JcSEP2*, and *JcSEP5* were upregulated in leaves, whereas *JcLhca3*, *JcLhca5*, *JcLhcb1.1*, *JcLhcb1.2*,

238 *JcLhcb3*, *JcLhcb5*, *JcLhcb6*, *JcLhcb8*, *JcSEP1*, *JcSEP3*, *JcSEP4*, and *JcPsb33* were downregulated; *JcLhca1*,

239 *JcLhcb1.1*, *JcLhcb1.2*, *JcLhcb3*, *JcLhcb5*, and *JcFCII* were downregulated in roots, whereas *JcOHP1* was

240 upregulated (**Fig. 5**).

241   Similar to drought treatment, after applying 100 mM NaCl for 2 h, 2 d or 7 d, gradual downregulation of

242   total transcripts was only observed in leaves. For 2 h, five or three genes were significantly regulated in leaves

243   and roots, respectively. Among them, *JcLhcb2*, *JcOHP2*, *JcSEP2*, and *JcFCII* were downregulated in leaves,

244   whereas *JcLhcb1.2* was upregulated; in roots, *JcLhcb8*, *JcSEP5*, and *JcFCII* were downregulated. For 2 d, six

245   genes were significantly regulated in both leaves and roots, respectively. Among them, *JcELIP*, *JcSEP2*,

246   *JcSEP4*, and *JcFCII* were upregulated in leaves, whereas *JcOHP2* and *JcSEP1* were downregulated; *JcLhca1*,

247   *JcLhca2*, *JcLhcb8*, and *JcSEP5* were upregulated in roots, whereas *JcLhcb1.1* and *JcLhcb1.2* were

248   downregulated. For 7 d, nine or seven genes were significantly regulated in both leaves and roots, respectively.

249   Among them, *JcLhca3*, *JcLhcb1.1*, *JcLhcb1.2*, *JcLhcb3*, *JcLhcb5*, *JcLhcb8*, *JcSEP1*, and *JcSEP4* were

250   downregulated in leaves, whereas *JcLhcb2* was upregulated; *JcLhca3*, *JcLhcb1.2*, *JcLhcb3*, *JcLhcb5*, and

251   *JcSEP5* were downregulated in roots, whereas *JcELIP* and *JcFCII* were upregulated (**Fig. 5**). Downregulation

252   of most regulated genes is highly consistent with gradual decrease of Pn, where the Pn values of 2 and 7 d after

253   stress treatment accounted for 83% or 50% of the control, respectively (*Zhang et al., 2014*).

254   Responses to gibberellin acid (GA) and 6-benzylaminopurine (BA) treatments were also examined in young

255   axillary buds. Application of 10 μM GA for 12 h resulted in one upregulated (i.e. *JcFCII*) and two

256   downregulated (i.e. *JcSEP2* and *JcSEP5*) genes, in contrast, no evident effect was observed for the same

257   concentration of BA (**Fig. 5**).

258   **DISCUSSION**

259   In green plants, the Lhc superfamily is consisted of four antennae protein families that play essential roles in

260   light-harvesting and photoprotection (*Jansson, 1999; Klimmek et al., 2006; Zou, 2018*). Despite their

261   importance, extensive research is still limited to the model plant arabidopsis and few other species such as

262   *Chlamydomonas reinhardtii*, *Physcomitrella patens*, castor, and cassava (*Elrad & Grossman, 2004; Klimmek*

263   *et al., 2006; Alboresi et al., 2008; Engelken et al., 2010; Zou et al., 2013; Zou & Yang, 2019a*). Among them,

264   it's well established that cassava and arabidopsis experienced one or two additional WGDs after the so-called γ

265   hexaploidization event shared by all core eudicots: The recent WGD occurred in cassava is called ρ, which was

266   estimated to occur within a window of 39–47 million years ago (Mya) (*Bredeson et al., 2016; Zou et al., 2019;*

267   *Zou & Yang, 2019a, 2019b, 2019c*), whereas two recent WGDs occurred in arabidopsis are known as β and α,

268   which were estimated to occur within a window of 61–65 or 23–50 Mya, respectively (*Bowers et al., 2003*).

269   From this point of view, analysis of species without recent WGDs may improve our knowledge on species-

270   specific evolution of this special gene family. Jatropha, another economically Euphorbiaceous plant for

271   potential biodiesel purpose, is a good candidate for such study. According to comparative genomics analysis,

272   jatropha may share a common ancestor with cassava and castor at approximately 65 Mya (*Bredeson et al.,*

273   *2016*), and no additional WGD occurred after their divergence.

274   In the present study, a first genome-wide identification and global analysis of *Lhc* superfamily genes were

275   performed in jatropha, and the superfamily number of 27 members identified in this species is comparable to

276   28 reported in castor but relatively less than 35 or 34 present in cassava and arabidopsis, respectively (*Klimmek*

277   *et al., 2006; Engelken et al., 2010; Zou et al., 2013; Zou & Yang, 2019a*). Nevertheless, 26 OGs representing

278   four families (i.e. Lhc, Lil, PsbS, and FCII) were found in jatropha as observed in castor and cassava (*Zou et*

279   *al., 2013; Zou & Yang, 2019a*). Few recent duplicates were identified in jatropha as well as castor,

280   corresponding to no recent WGD occurred in these two species (*Chan et al., 2010; Wu et al., 2015*). Compared

281   with castor that contains two dispersed duplicates, only one duplicate derived from tandem duplication was

282   found in jatropha. By contrast, considerably more duplicates, i.e. nine, were identified in both cassava and

283   arabidopsis (**Supplementary Table S1**), reflecting the occurrence of one or two recent WGDs (*Bowers et al.,*

284   *2003; Bredeson et al., 2016*). Despite having the same number of duplicates, cassava contains relatively more

285   WGD duplicates (6 vs 5) but less local duplicates (3 vs 4), implying species-specific evolution pattern.

286   Interestingly, duplicates in both jatropha and castor are confined to the Lhcb1 group, however, in cassava and

287   arabidopsis, local duplicates were also found in the Lhcb2 group (**Supplementary Table S1**).

288   Among four families identified, both PsbS and FCII include a single member in four species examined in

289   this study. By contrast, Lhc and Lil families are relatively complex. The Lhc family contains 14 OGs

290   representing two subfamilies (i.e. Lhca and Lhcb), whereas the Lil family includes ten OGs representing four

291   subfamilies (i.e. ELIP, OHP, SEP, and Psb33). According to crystal analyses, the Lhc family members usually

292   contain three alpha-helixes, whereas the PsbS family features four (*Kühlbrandt et al., 1994; Pan et al., 2011;*

293   *Fan et al., 2015*). By contrast, one to three helixes were shown to be present in Lil proteins, i.e. one for OHPs

294   and Psb33s, two for SEPs, and three for ELIPs (*Jansson, 1999; Engelken et al., 2010; Fristedt et al., 2015;*

295   *Beck et al., 2017*). Similar results were also observed in this study, however, both JcSEP5 and JcFCII were

296   shown to contain a single helix.

297   It is noteworthy that, among 26 OGs identified, SEP6 is absent from arabidopsis. SEP6 exhibits about

298   41.7%, 40.5% or 39.0–40.3% sequence identity with SEP3 in jatropha, castor and cassava respectively,

299   implying their early divergence and species-specific gene loss. Indeed, SEP6 orthologs were broadly found in

300   dicots, including *Carica papaya* and *Aquilegia coerulea* (*Zou & Yang, 2019a*). Additionally, Lhcb8 shows

301   approximately 72.0%, 69.9%, 72.6% or 64.2–65.3% identity with Lhcb4 in jatropha, castor, cassava, and

302   arabidopsis, respectively. Lhcb8 is widely present in core eudicots but not in *A. coerulea* and monocots,

303   suggesting its recent origin. According to synteny analysis performed in arabidopsis, *Lhcb8* is more likely to

304   be a duplicate of *Lhcb4* generated along the γ event (*Bowers et al., 2003; Wang et al., 2013*).

305   Potential roles of *JcLhc* superfamily genes could be inferred from their expression patterns and function-

306   characterized orthologs in arabidopsis and other species. According to GO annotation, they belong to thylakoid

307   membrane proteins that have activity of chlorophyll binding, pigment binding, xanthophyll binding, lipid

308   binding, protein binding, iron-sulfur cluster binding, oxidoreductase, ferrochelatase as summarized in

309   **Supplementary Table S4**. Our transcriptional profiling not only supports the expression of all 27 *JcLhc*

310   superfamily genes identified in this study, but also reveals key genes in a certain tissue, development stage or

311   environment condition. Similar to that reported in arabidopsis (*Jansson, 1999; Klimmek et al., 2006*), genes

312   encoding JcLhca1 to -4 and JcLhcb1 to -6, which are characterized as abundant Lhc proteins, were highly

313   expressed in most examined jatropha tissues, especially in mature leaf. By contrast, four genes encoding so-

314   called rare Lhc proteins (i.e. *JcLhca5*, *JcLhca6*, *JcLhcb7*, and *JcLhcb8*) are lowly expressed, exhibiting a

315   similar expression pattern to members of Lil, PsbS, and FCII families. The result is consistent with our cluster

316   analysis, which divides *JcLhc* superfamily genes into five clusters named I, II, III, IV, and V. These genes

317   encoding abundant Lhc proteins belong to Clusters II and V, whereas rare *Lhc* genes were divided into Cluster

318   III with the exception of *JcLhca5*. Clustering *JcLhca5* into Cluster II is due to its leaf-preferential expression

319   pattern, but not the transcript level. Compared with mature leaf, leafage is considerably more sensitive to high

320   light and other stresses. This is not surprising that the majority of members in Lil, PsbS, and FCII families are

321   highly expressed in this special tissue, comprising Clusters I and III. It is worth noting that, *JcSEP6*, the unique

322   member in Clusters IV, is lowly expressed in most examined tissues with the exception of root. As a recently

323   identified superfamily member, the detailed function of SEP6 still needs to be investigated. Furthermore, most

324   *JcLhc* superfamily genes were regulated by drought and/or salt, two most important abiotic stresses affecting

325   crop growth and yield (*Zhang et al., 2014, 2015*). As expected, more genes are downregulated, especially for

326   the leaf tissue, corresponding to decrease of transpiration rate and stomatal conductance (*Zhang et al., 2015*).

327   Nevertheless, frequent upregulation of certain members was also observed, i.e. *JcLhca4*, *JcELIP*, *JcPsbS*,

328   *JcSEP2*, *JcSEP5*, and *JcOHP1*. Interestingly, *JcLhca2* and *JcLhca5* exhibit distinct responses upon drought or

329   salt stress, whereas *JcLhca2* was specifically regulated by salt and *JcLhca4*, *JcLhcb4*, *JcLhcb6*, *JcLhcb7*,

330   *JcPsbS*, *JcOHP1*, *JcSEP2* and *JcPsb33* were only regulated by drought. The involvement of *Lhc* superfamily

331   genes in stress response has been well documented in arabidopsis and other species, including high light,

332   chloroplast retrograde signal, oxidative stress, abscisic acid, etc. (*Montané & Kloppstech, 2000; Staneloni et*

333   *al., 2008; Gerotto et al., 2011; Tibiletti et al., 2016*). For example, ELIPs, as the name suggests, are induced

334   by early and high light, as well as other stresses such as UV-B, cold, heat, drought, salt, hypoxia, and anoxia

335   described in this study and elsewhere (*Heddad et al., 2006; Hayami et al., 2015*). In arabidopsis, knockout and

336   overexpression of *ELIPs* resulted in decreased chlorophyll levels (*Casazza et al., 2005; Tzvetkova-Chevolleau*

337   *et al., 2007*). Another high-light induced gene, *PsbS*, acts as the main sensor of the low pH in plants and plays

338   an essential role in nonphotochemical quenching (*Bergantino et al., 2003; Li et al., 2004; Gerotto et al., 2011;*

339   *Fan et al., 2015; Tibiletti et al., 2016*). OHPs, which are related to HLIPs in cyanobacteria, are essential for the

340   formation of the PSII reaction center. In arabidopsis, mutations in *AtOHP1* or *AtOHP2* caused severe growth

341   deficits, reduced pigmentation, and disturbed thylakoid architecture (*Beck et al., 2017; Hey and Grimm, 2018;*

342   *Myouga et al., 2018; Li et al., 2019*). By contrast, two plant hormones (i.e. BA and GA), which can improve

343   shoot branching after application to young axillary buds (*Ni et al., 2017*), had little effect on transcriptional

344   regulation of *JcLhc* superfamily genes.

345   **CONCLUSION**

346   This study presents a first genomics analysis of the *Lhc* supergene family in jatropha, resulting in 27 members

347   that are distributed across nine out of 11 chromosomes. Despite a relatively smaller number of members, 26

348   orthologous groups representing four families were found, where SEP6 represents a novel group that has been

349   lost in the model plant arabidopsis. Nearly one-to-one orthologous relationship was observed between jatropha

350   and castor, however, species-specific gene expansion was observed in these two species as well as cassava and

351   arabidopsis. Exon-intron structures, protein motifs, and expression profiles of *JcLhc* superfamily genes were

352   also analyzed and discussed. These findings provide valuable information for further studies in jatropha and

353   species beyond.

354   **ACKNOWLEDGEMENTS**

355   The authors appreciate those contributors who make the related genome and transcriptome data accessible in

356   public databases. They also thank the editor and two reviewers for their helpful suggestions.

357   **ADDITIONAL INFORMATION AND DECLARATIONS**

358   **Funding**

359   This work was supported by the National Natural Science Foundation of China (31700580 and 31971688), the

360   Natural Science Foundation of Hainan province (319MS093), the Central Public-interest Scientific Institution

361   Basal Research Fund for Chinese Academy of Tropical Agricultural Sciences (1630022019017 and

362   1630052017011), and the Research Fund of Guangdong University of Petrochemical Technology (2018rc55).

363   The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the

364   manuscript.

365   **Grant Disclosures**

366   The following grant information was disclosed by the authors:

367   National Natural Science Foundation of China: 31700580 and 31971688.

368   Natural Science Foundation of Hainan province: 319MS093.

369   Central Public-interest Scientific Institution Basal Research Fund for Chinese Academy of

370   Tropical Agricultural Sciences: 1630052017011 and 1630022019017.

371   Research Fund of Guangdong University of Petrochemical Technology: 2018rc55.

372   **Competing interests**

373   The authors declare there are no competing interests.

374   **Supplemental Information**

375   Supplemental information for this article can be found online at http://dx.doi.org/.

376 **REFERENCES**

377 **Alboresi A, Caffarri S, Nogue F, Bassi R, Morosinotto T. 2008.** *In silico* and biochemical analysis of
378   *Physcomitrella patens* photosynthetic antenna: identification of subunits which evolved upon land adaptation.
379   *PLoS One* **3(4)**:e2033 DOI 10.1371/journal.pone.0002033.

380 **Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997.** Gapped BLAST and
381   PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25(17)**:3389–3402
382   DOI 10.1093/nar/25.17.3389.

383 **Beck J, Lohscheider JN, Albert S, Andersson U, Mendgen KW, Rojas-Stütz MC, Adamska I, Funck D. 2017.**
384   Small one-helix proteins are essential for photosynthesis in *Arabidopsis*. *Frontiers in Plant Science* **8**:7 DOI
385   10.3389/fpls.2017.00007.

386 **Berchmans HJ, Hirata S. 2008.** Biodiesel production from crude *Jatropha curcas* L. seed oil with a high content of
387   free fatty acids. *Bioresource Technology* **99(6)**:1716–1721 DOI 10.1016/j.biortech.2007.03.051.

388 **Bergantino E, Segalla A, Brunetta A, Teardo E, Rigoni F, Giacometti GM, Szabò I. 2003.** Light- and pH-
389   dependent structural changes in the PsbS subunit of photosystem II. *Proceedings of the National Academy of*
390   *Sciences of the United States of America* **100(25)**:15265–15270 DOI 10.1073/pnas.2533072100.

391 **Bowers JE, Chapman BA, Rong J, Paterson AH. 2003.** Unravelling angiosperm genome evolution by
392   phylogenetic analysis of chromosomal duplication events. *Nature* **422(6930)**:433–438 DOI
393   10.1038/nature01521.

394 **Bredeson JV, Lyons JB, Prochnik SE, Wu GA, Ha CM, Edsinger-Gonzales E, Grimwood J, Schmutz J,**
395   **Rabbi IY, Egesi C, Nauluvula P, Lebot V, Ndunguru J, Mkamilo G, Bart RS, Setter TL, Gleadow RM,**
396   **Kulakow P, Ferguson ME, Rounsley S, Rokhsar DS. 2016.** Sequencing wild and cultivated cassava and
397   related species reveals extensive interspecific hybridization and genetic diversity. *Nature Biotechnology*
398   **34(5)**:562–70. DOI 10.1038/nbt.3535.

399 **Cai B, Yang X, Tuskan GA, Cheng ZM. 2011.** MicroSyn: a user friendly tool for detection of microsynteny in a
400   gene family. *BMC Bioinformatics* **12**:79 DOI 10.1186/1471-2105-12-79.

401 **Casazza AP, Rossini S, Rosso MG, Soave C. 2005.** Mutational and expression analysis of ELIP1 and ELIP2 in
402   *Arabidopsis thaliana*. *Plant Molecular Biology* **58(1)**:41–51 DOI 10.1007/s11103-005-4090-1.

403 **Chan AP, Crabtree J, Zhao Q, Lorenzi H, Orvis J, Puiu D, Melake-Berhan A, Jones KM, Redman J, Chen G,**
404   **Cahoon EB, Gedil M, Stanke M, Haas BJ, Wortman JR, Fraser-Liggett CM, Ravel J, Rabinowicz PD.**
405   **2010.** Draft genome sequence of the oilseed species *Ricinus communis*. *Nature Biotechnology* **28(9)**:951–956
406   DOI 10.1038/nbt.1674.

407 **Edgar RC. 2004.** MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids*
408   *Research* **32**:1792–1797 DOI 10.1093/nar/gkh340.

409 **Engelken J, Brinkmann H, Adamska I. 2010.** Taxonomic distribution and origins of the extended LHC (light-
410   harvesting complex) antenna protein superfamily. *BMC Evolution Biology* **10**:233 DOI 10.1186/1471-2148-10-

411   233.

412 **Elrad D, Grossman AR. 2004.** A genome's-eye view of the light-harvesting polypeptides of *Chlamydomonas*
413   *reinhardtii*. *Current Genetics* **45(2)**:61–75 DOI 10.1007/s00294-003-0460-x.

414 **Fairless D. 2007.** Biofuel: the little shrub that could--maybe. *Nature* **449(7163)**:652–655 DOI 10.1038/449652a.

415 **Fan M, Li M, Liu Z, Cao P, Pan X, Zhang H, Zhao X, Zhang J, Chang W. 2015.** Crystal structures of the PsbS
416   protein essential for photoprotection in plants. *Nature Structural & Molecular Biology* **22(9)**:729–735 DOI
417   10.1038/nsmb.3068.

418 **Fristedt R, Herdean A, Blaby-Haas CE, Mamedov F, Merchant SS, Last RL, Lundin B. 2015.**
419   PHOTOSYSTEM II PROTEIN33, a protein conserved in the plastid lineage, is associated with the chloroplast
420   thylakoid membrane and provides stability to photosystem II supercomplexes in *Arabidopsis*. *Plant Physiology*
421   **167(2)**:481–492 DOI 10.1104/pp.114.253336.

422 **Gerotto C, Alboresi A, Giacometti GM, Bassi R, Morosinotto T. 2011.** Role of PSBS and LHCSR in
423   *Physcomitrella patens* acclimation to high light and low temperature. *Plant Cell & Environment* **34**:922–932
424   DOI 10.1111/j.1365-3040.2011.02294.x.

425 **Heddad M, Norén H, Reiser V, Dunaeva M, Andersson B, Adamska I. 2006.** Differential expression and
426   localization of early light-induced proteins in *Arabidopsis*. *Plant Physiology* **142(1)**:75–87 DOI
427   10.1104/pp.106.081489.

428 **Hayami N, Sakai Y, Kimura M, Saito T, Tokizawa M, Iuchi S, Kurihara Y, Matsui M, Nomoto M, Tada Y,**
429   **Yamamoto YY. 2015.** The responses of *Arabidopsis* early light-induced protein2 to ultraviolet B, high light,
430   and cold stress are regulated by a transcriptional regulatory unit composed of two elements. *Plant Physiology*
431   **169(1)**:840–855 DOI 10.1104/pp.15.00398.

432 **Hey D, Grimm B. 2018.** Requirement of ONE-HELIX PROTEIN 1 (OHP1) in early *Arabidopsis* seedling
433   development and under high light intensity. *Plant Signaling & Behavior* **29**:1–3 DOI
434   10.1080/15592324.2018.1550317.

435 **Hey D, Rothbart M, Herbst J, Wang P, Müller J, Wittmann D, Gruhl K, Grimm B. 2017.** LIL3, a light-
436   harvesting complex protein, links terpenoid and tetrapyrrole biosynthesis in *Arabidopsis thaliana*. *Plant*
437   *Physiology* **174(2)**:1037–1050 DOI 10.1104/pp.17.00505.

438 **Hobe S, Förster R, Klingler J, Paulsen H. 1995.** N-proximal sequence motif in light-harvesting chlorophyll a/b-
439   binding protein is essential for the trimerization of light-harvesting chlorophyll a/b complex. *Biochemistry*
440   **34(32)**:10224–10228.

441 **Jansson S. 1999.** A guide to the Lhc genes and their relatives in *Arabidopsis*. *Trends in Plant Science* **4(6)**:236–240
442   DOI 10.1016/S1360-1385(99)01419-3.

443 **Klimmek F, Sjödin A, Noutsos C, Leister D, Jansson S. 2006.** Abundantly and rarely expressed Lhc protein genes
444   exhibit distinct regulation patterns in plants. *Plant Physiology* **140(3)**:793–804 DOI 10.1104/pp.105.073304.

445 **Koziol AG, Borza T, Ishida K, Keeling P, Lee RW, Durnford DG. 2007.** Tracing the evolution of the light-
446   harvesting antennae in chlorophyll a/b-containing organisms. *Plant Physiology* **143(4)**:1802–1816 DOI

447   10.1104/pp.106.092536.

448   **Kumar A, Sharma S. 2008.** An evaluation of a multipurpose oil seed crop for industrial uses (*Jatropha curcas* L): a
449   review. *Industrial Crops & Products* **28**:1–10 DOI 10.1016/j.indcrop.2008.01.001.

450   **Kumar S, Stecher G, Tamura K. 2016.** MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger
451   datasets. *Molecular Biology and Evolution* **33(7)**:1870–1874 DOI 10.1093/molbev/msw054.

452   **Kühlbrandt W, Wang DN, Fujiyoshi Y. 1994.** Atomic model of plant light-harvesting complex by electron
453   crystallography. *Nature* **367(6464)**:614–621 DOI 10.1038/367614a0.

454   **Langmead B, Salzberg SL. 2012.** Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9(4)**:357–359 DOI
455   10.1038/nmeth.1923.

456   **Li B, Dewey CN. 2011.** RSEM: accurate transcript quantification from RNA-Seq data with or without a reference
457   genome. *BMC Bioinformatics***12**:323 DOI 10.1186/1471-2105-12-323.

458   **Li H, Tsuchimoto S, Harada K, Yamasaki M, Sakai H, Wada N, Alipour A, Sasai T, Tsunekawa A, Tsujimoto**
459   **H, Ando T, Tomemori H, Sato S, Hirakawa H, Quintero VP, Zamarripa A, Santos P, Hegazy A, Ali AM,**
460   **Fukui K. 2017.** Genetic tracing of *Jatropha curcas* L. from its mesoamerican origin to the world. *Frontiers in*
461   *Plant Science* **8**:1539 DOI 10.3389/fpls.2017.01539.

462   **Li XP, Gilmore AM, Caffarri S, Bassi R, Golan T, Kramer D, Niyogi KK. 2004.** Regulation of photosynthetic
463   light harvesting involves intrathylakoid lumen pH sensing by the PsbS protein. *Journal of Biological Chemistry*
464   **279**:22866–22874 DOI 10.1074/jbc.M402461200.

465   **Li Y, Liu B, Zhang J, Kong F, Zhang L, Meng H, Li W, Rochaix JD, Li D, Peng L. 2019.** OHP1, OHP2, and
466   HCF244 form a transient functional complex with the photosystem II reaction center. *Plant Physiology*
467   **179(1)**:195–208 DOI 10.1104/pp.18.01231.

468   **Maghuly F, Laimer M. 2013.** *Jatropha curcas*, a biofuel crop: functional genomics for understanding metabolic
469   pathways and genetic improvement. *Biotechnology Journal* **8**:1172–1182 DOI 10.1002/biot.201300231.

470   **Mazumdar P, Singh P, Babu S, Siva R, Harikrishna JA. 2018.** An update on biological advancement of *Jatropha*
471   *curcas* L.: New insight and challenges. *Renewable & Sustainable Energy Reviews* **91**:903–917 DOI
472   10.1016/j.rser.2018.04.082.

473   **Montané MH, Kloppstech K. 2000.** The family of light-harvesting-related proteins (LHCs, ELIPs, HLIPs): was the
474   harvesting of light their primary function? *Gene* **258(1-2)**:1–8 DOI 10.1016/S0378-1119(00)00413-3.

475   **Montes JM, Melchinger AE. 2016.** Domestication and breeding of *Jatropha curcas* L. *Trends in Plant Science*
476   **21**:1045–1057 DOI 10.1016/j.tplants.2016.08.008.

477   **Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008.** Mapping and quantifying mammalian
478   transcriptomes by RNA-Seq. *Nature Methods* **5(7)**:621–628 DOI 10.1038/nmeth.1226.

479   **Myouga F, Takahashi K, Tanaka R, Nagata N, Kiss AZ, Funk C, Nomura Y, Nakagami H, Jansson S,**
480   **Shinozaki K. 2018.** Stable accumulation of photosystem II requires ONE-HELIX PROTEIN1 (OHP1) of the
481   light harvesting-like family. *Plant Physiology* **176(3)**:2277–2291 DOI 10.1104/pp.17.01782.

482   **Ni J, Zhao ML, Chen MS, Pan BZ, Tao YB, Xu ZF. 2017.** Comparative transcriptome analysis of axillary buds in

483   response to the shoot branching regulators gibberellin A3 and 6-benzyladenine in *Jatropha curcas*. *Scientific*
484   *Reports* **7(1)**:11417 DOI 10.1038/s41598-017-11588-0.

485   **Pan X, Li M, Wan T, Wang L, Jia C, Hou Z, Zhao X, Zhang J, Chang W. 2011.** Structural insights into energy
486   regulation of light-harvesting complex CP29 from spinach. *Nature Structural & Molecular Biology* **18(3)**:309–
487   315 DOI 10.1038/nsmb.2008.

488   **Rogl H, Kühlbrandt W. 1999.** Mutant trimers of light-harvesting complex II exhibit altered pigment content and
489   spectroscopic features. *Biochemistry* **38(49)**:16214–16222 DOI 10.1021/bi990739p.

490   **Staneloni RJ, Rodriguez-Batiller MJ, Casal JJ. 2008.** Abscisic acid, high-light, and oxidative stress down-
491   regulate a photosynthetic gene via a promoter motif not involved in phytochrome-mediated transcriptional
492   regulation. *Molecular Plant* **1(1)**:75–83 DOI 10.1093/mp/ssm007.

493   **Tibiletti T, Auroy P, Peltier G, Caffarri S. 2016.** *Chlamydomonas reinhardtii* PsbS protein is functional and
494   accumulates rapidly and transiently under high light. *Plant Physiology* **171(4)**:2717–2730 DOI
495   10.1104/pp.16.00572.

496   **Tzvetkova-Chevolleau T, Franck F, Alawady AE, Dall'Osto L, Carrière F, Bassi R, Grimm B, Nussaume L,**
497   **Havaux M. 2007.** The light stress-induced protein ELIP2 is a regulator of chlorophyll synthesis in *Arabidopsis*
498   *thaliana*. *Plant Journal* **50(5)**:795–809 DOI 10.1111/j.1365-313X.2007.03090.x.

499   **Voorrips RE. 2002.** MapChart: software for the graphical presentation of linkage maps and QTLs. *Journal of*
500   *Heredity* **93**:77–78.

501   **Wang Y, Tan X, Paterson AH. 2013.** Different patterns of gene structure divergence following gene duplication in
502   *Arabidopsis*. *BMC Genomics* **14**:652 DOI 10.1186/1471-2164-14-652.

503   **Wu P, Zhou C, Cheng S, Wu Z, Lu W, Han J, Chen Y, Chen Y, Ni P, Wang Y, Xu X, Huang Y, Song C,**
504   **Wang Z, Shi N, Zhang X, Fang X, Yang Q, Jiang H, Chen Y, Li M, Wang Y, Chen F, Wang J, Wu G.**
505   **2015.** Integrated genome sequence and linkage map of physic nut (*Jatropha curcas* L.), a biodiesel plant. *Plant*
506   *Journal* **81(5)**:810–821 DOI 10.1111/tpj.12761.

507   **Zhang C, Zhang L, Zhang S, Zhu S, Wu P, Chen Y, Li M, Jiang H, Wu G. 2015.** Global analysis of gene
508   expression profiles in physic nut (*Jatropha curcas* L.) seedlings exposed to drought stress. *BMC Plant Biology*
509   **15**:17 DOI 10.1186/s12870-014-0397-x.

510   **Zhang L, Zhang C, Wu P, Chen Y, Li M, Jiang H, Wu G. 2014.** Global analysis of gene expression profiles in
511   physic nut (*Jatropha curcas* L.) seedlings exposed to salt stress. *PLoS One* **9(5)**:e97878 DOI
512   10.1371/journal.pone.0097878.

513   **Zou Z. 2018.** Mining gene families in the castor bean genome. In: Rabinowicz P, Kole C (Eds.): The Castor Bean
514   Genome, Springer, Switzerland, pp. 135–173.

515   **Zou Z, Huang QX, An F. 2013.** Genome-wide identification, classification and expression analysis of *Lhc*
516   supergene family in castor bean (*Ricinus communis* L.). *Agriculture Biotechnology* **2**:44–48, 51.

517   **Zou Z, Huang QX, Xie GS, Yang LF. 2018.** Genome-wide comparative analysis of papain-like cysteine protease
518   family genes in castor bean and physic nut. *Scientific Reports* **8**:331 DOI 10.1038/s41598-017-18760-6.

Zou Z, Yang JH. 2019a. Genomics analysis of the light-harvesting chlorophyll a/b-binding (Lhc) superfamily in cassava (*Manihot esculenta* Crantz). *Gene* **702**:171–181 DOI 10.1016/j.gene.2019.03.071.

Zou Z, Yang JH. 2019b. Genome-wide comparison reveals divergence of cassava and rubber aquaporin family genes after the recent whole-genome duplication. *BMC genomics* **20**:380 DOI 10.1186/s12864-019-5780-4.

Zou Z, Yang JH. 2019c. Genomic analysis of Dof transcription factors in *Hevea brasiliensis*, a rubber-producing tree. *Industrial Crops & Products* **134**:271–283 DOI 10.1016/j.indcrop.2019.04.013.

Zou Z, Yang JH, Zhang XC. 2019. Insights into genes encoding respiratory burst oxidase homologs (RBOHs) in rubber tree (*Hevea brasiliensis* Muell. Arg.). *Industrial Crops & Products* **128**:126–139 DOI 10.1016/j.indcrop.2018.11.005.

Zou Z, Yang LF, Gong J, Mo YY, Wang JK, Cao JH, An F, Xie GS. 2016. Genome-wide identification of *Jatropha curcas* aquaporin genes and the comparative analysis provides insights into the gene family expansion and evolution in *Hevea brasiliensis*. *Frontiers in Plant Science* **7**:395 DOI 10.3389/fpls.2016.00395.

Figure and Table legends

Fig. 1 Chromosomal locations of *JcLhc* superfamily genes. Chromosome serial numbers are indicated at the top of each chromosome. (Chr: chromosome)

Fig. 2 Exon-intron structures of *JcLhc* superfamily genes. The graphic representation of the gene models is displayed using GSDS. (GSDS: gene structure display server)

Fig. 3 Phylogenetic and conserved motif analyses of jatropha, castor, cassava, and arabidopsis Lhc superfamily proteins. a. Phylogenetic analysis of Lhca, Lhcb, PsbS, ELIP, OHP, SEP, Psb33, and FCII subfamilies; **b.** Distribution of conserved motifs. Sequence alignment was performed using MUSCLE and unrooted phylogenetic trees were constructed using MEGA7 (maximum likelihood method; bootstrap, 1,000 replicates). Only bootstrap values at nodes supported by a posterior probability of ≥50% are given. The distance scale denotes the number of amino acid substitutions per site. The name of each OG is indicated next to the corresponding group. (OG: orthologous group)

Fig. 4 Tissue-specific expression profiles of *JcLhc* superfamily genes. Color scale represents FPKM normalized $\log_{10}$ transformed counts where navy indicates low expression and firebrick3 indicates high expression. (FPKM: fragments per kilobase of exon per million fragments mapped; IND: undifferentiated inflorescence of 0.5 cm diameter; PID1: female flower with carpel primordia beginning to differentiate; PID2:

547   female flower with three distinct carpels formed; STD1: male flower with stamen primordia beginning to

548   differentiate; STD2: male flower with ten complete stamens formed)

549   **Fig. 5 Expression profiles of *JcLhc* superfamily genes upon drought, salt, BA, or GA treatments.** Color

550   scale represents RPKM normalized $\log_{10}$ transformed counts where navy indicates low expression and

551   firebrick3 indicates high expression. (RPKM: Reads per kilobase per million mapped reads)

552   **Table 1 27 *Lhc* superfamily genes identified in jatropha.** (AA: amino acid; At: *Arabidopsis thaliana*; bp:

553   base pair; CDS: coding sequence; EST: expressed sequence tag; Me: *Manihot esculenta*; OG: orthologous

554   group; Rc: *Ricinus communis*; TMH: transmembrane helix; TP: transit peptide)

555   **Table 2 Detailed information of 25 motifs identified in this study.** Motifs were identified using MEME.

556

Figure 1

Fig. 1 Chromosomal locations of *JcLhc* superfamily genes.

Chromosome serial numbers are indicated at the top of each chromosome. (Chr: chromosome)
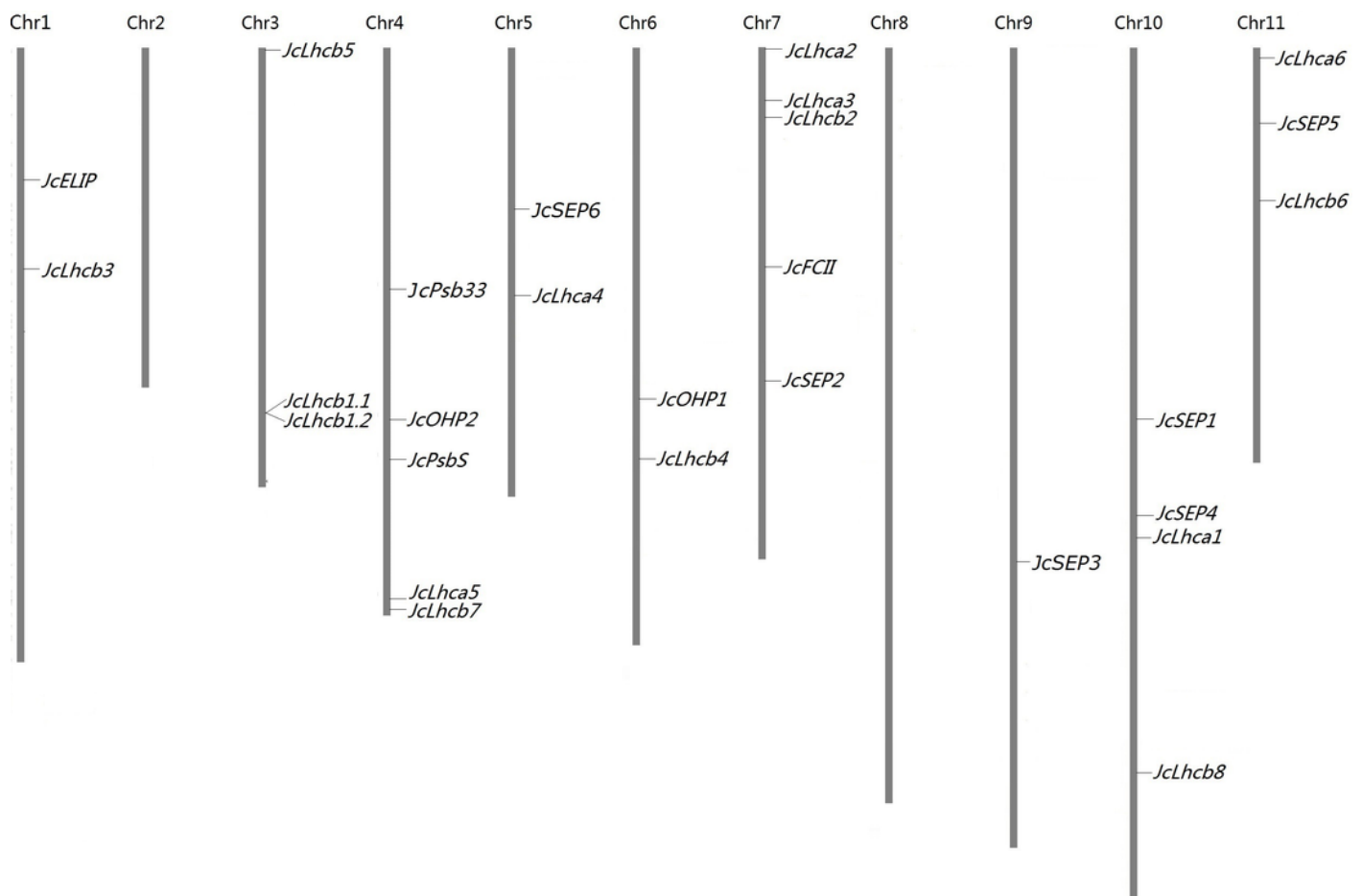
Figure 2

Fig. 2 Exon-intron structures of *JcLhc* superfamily genes.

The graphic representation of the gene models is displayed using GSDS. (GSDS: gene structure display server)
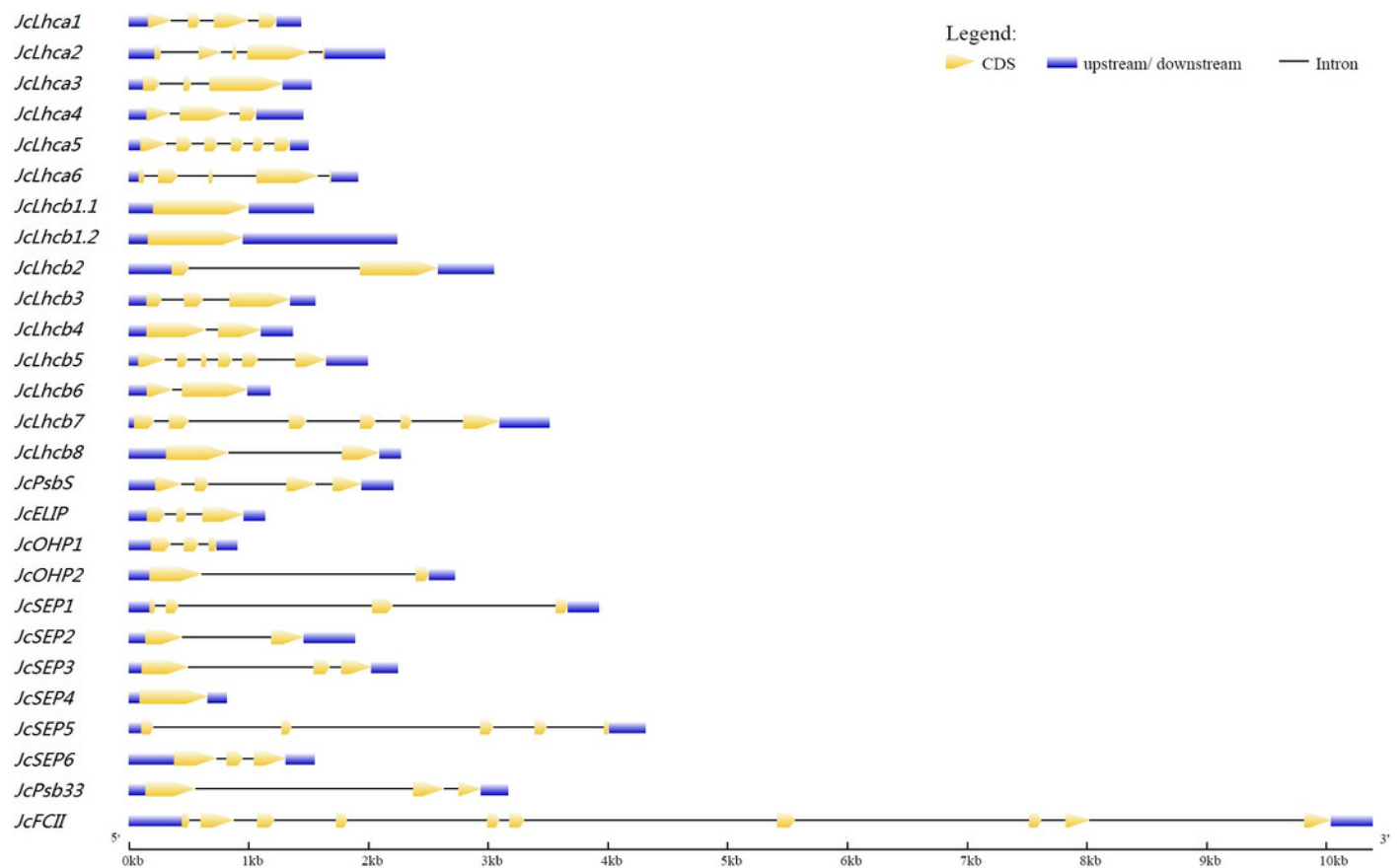
Figure 3

Fig. 3 Phylogenetic and conserved motif analyses of jatropha, castor, cassava, and arabidopsis Lhc superfamily proteins.

a. Phylogenetic analysis of Lhca, Lhcb, PsbS, ELIP, OHP, SEP, Psb33, and FCII subfamilies; **b.** Distribution of conserved motifs. Sequence alignment was performed using MUSCLE and unrooted phylogenetic trees were constructed using MEGA7 (maximum likelihood method; bootstrap, 1,000 replicates). Only bootstrap values at nodes supported by a posterior probability of ≥50% are given. The distance scale denotes the number of amino acid substitutions per site. The name of each OG is indicated next to the corresponding group. (OG: orthologous group)
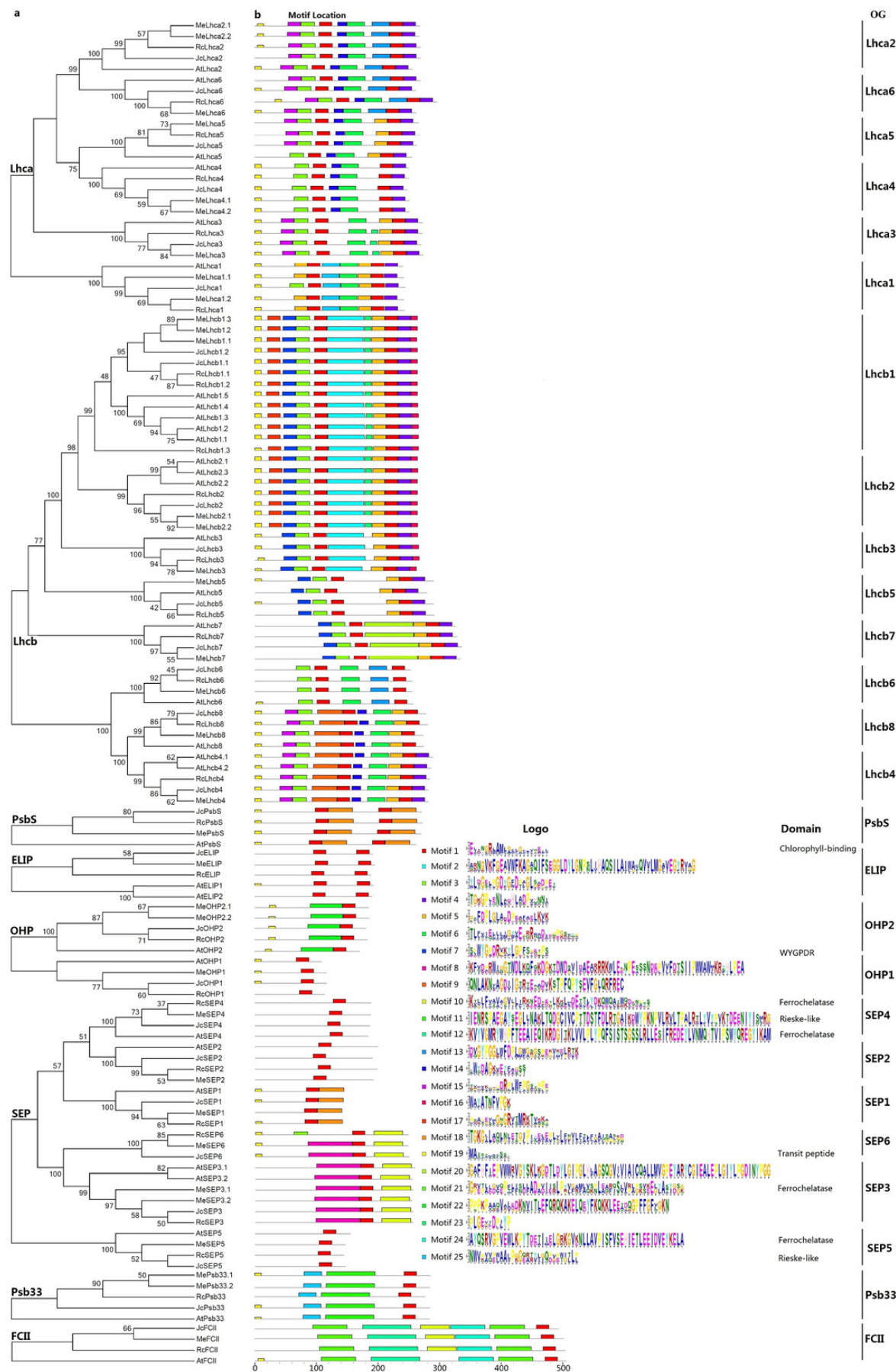
Figure 4

Fig. 4 Tissue-specific expression profiles of *JcLhc* superfamily genes.

Color scale represents FPKM normalized log_{10} transformed counts where navy indicates low expression and firebrick3 indicates high expression. (FPKM: fragments per kilobase of exon per million fragments mapped; IND: undifferentiated inflorescence of 0.5 cm diameter; PID1: female flower with carpel primordia beginning to differentiate; PID2: female flower with three distinct carpels formed; STD1: male flower with stamen primordia beginning to differentiate; STD2: male flower with ten complete stamens formed)
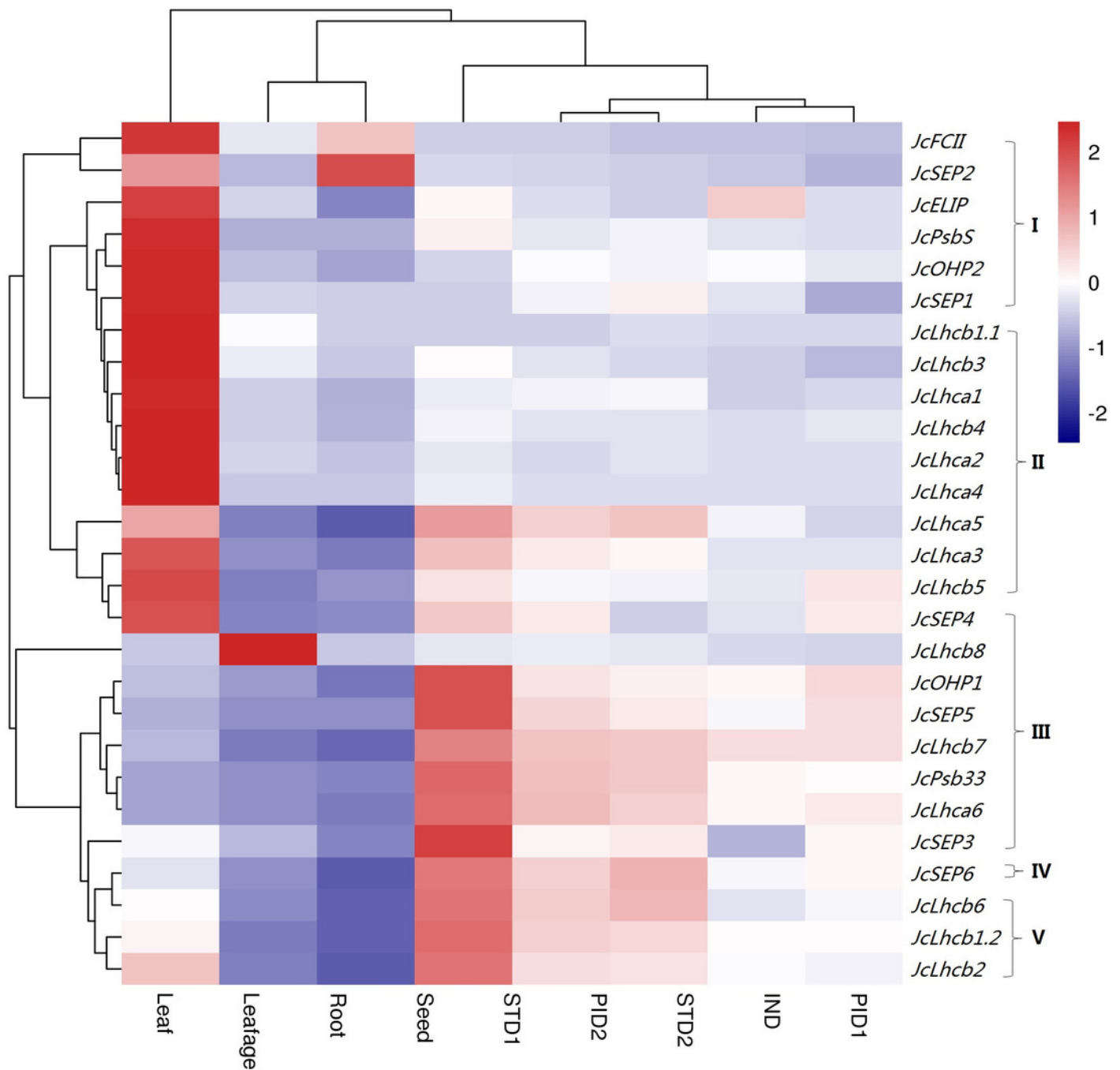
Figure 5

Fig. 5 Expression profiles of *JcLhc* superfamily genes upon drought, salt, BA, or GA treatments.

Color scale represents RPKM normalized $\log_{10}$ transformed counts where navy indicates low expression and firebrick3 indicates high expression. (RPKM: Reads per kilobase per million mapped reads)
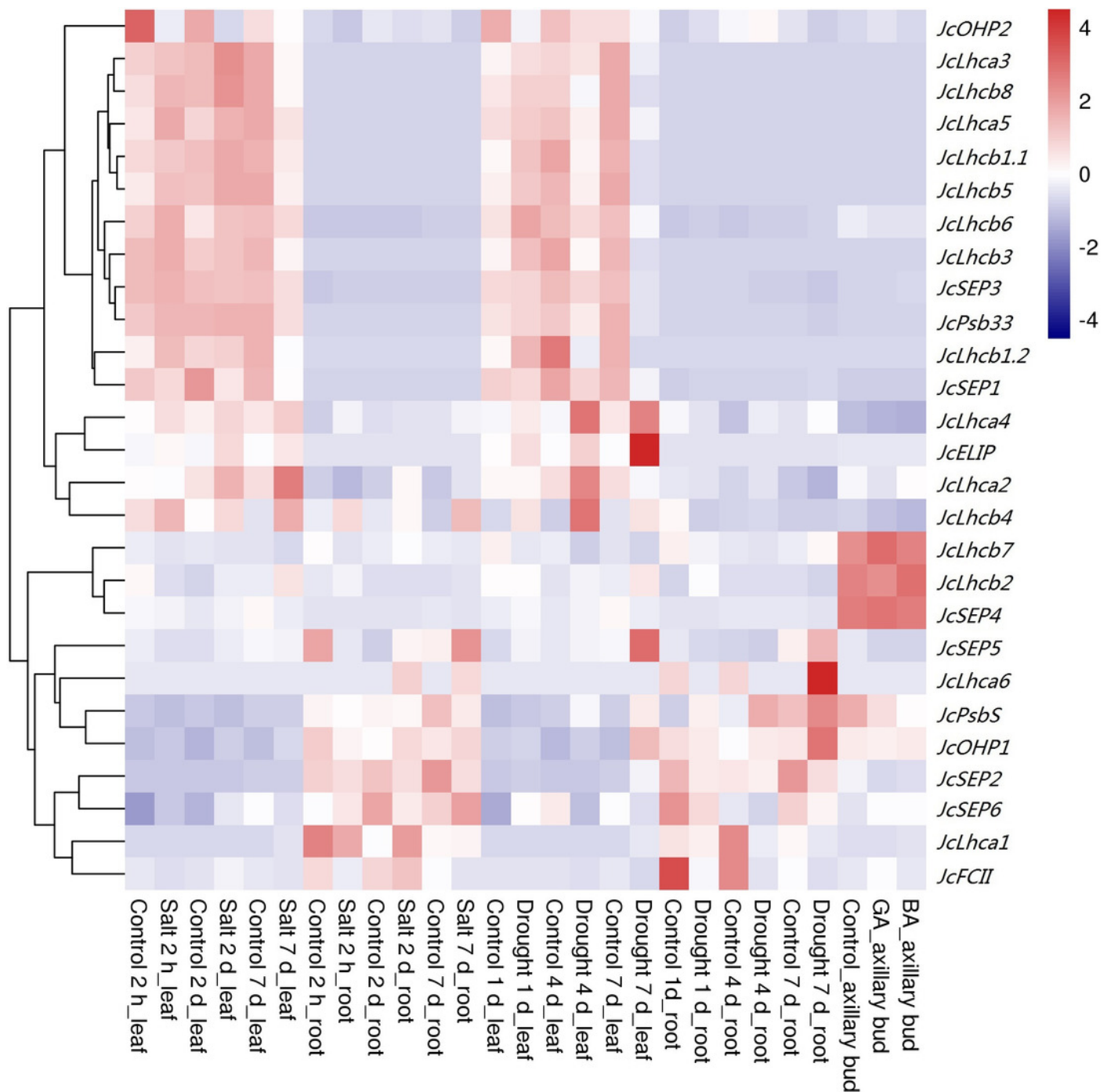
Table 1 27 *Lhc* superfamily genes identified in jatropha.

(AA: amino acid; At: *Arabidopsis thaliana*; bp: base pair; CDS: coding sequence; EST: expressed sequence tag; Me: *Manihot esculenta*; OG: orthologous group; Rc: *Ricinus communis*; TMH: transmembrane helix; TP: transit peptide)

Subfamily	Gene name	Locus ID	Scaffold position	Nucleotide length (bp, from start to stop codons) CDS	Gene	Intron no.	EST no.	AA	TP length	TMH	Ortholog Rc	Me	At	OG
Lhca	JcLhca1	JCGZ_23938	scaffold794:74368-75803(-)	738	1076	3	0	245	45	3	RcLhca1	MeLhca1.1 MeLhca1.2	AtLhca1	Lhca1
	JcLhca2	JCGZ_17961	scaffold502:3176588-3178724(+)	813	1491	4	1	270	58	3	RcLhca2	MeLhca2.1 MeLhca2.2	AtLhca2	Lhca2
	JcLhca3	JCGZ_15032	scaffold42:178242-179765(-)	816	1171	2	8	271	38	3	RcLhca3	MeLhca3	AtLhca3	Lhca3
	JcLhca4	JCGZ_11643	scaffold328:2197266-2195814(-)	750	917	2	2	249	48	3	RcLhca4	MeLhca4.1 MeLhca4.2	AtLhca4	Lhca4
	JcLhca5	JCGZ_04265	scaffold159:84206-85704(+)	795	1255	5	0	264	57	3	RcLhca5	MeLhca5	AtLhca5	Lhca5
	JcLhca6	JCGZ_07509	scaffold211:3089530-3091440(-)	792	1612	4	0	263	42	3	RcLhca6	MeLhca6	AtLhca6	Lhca6
Lhcb	JcLhcb1.1	JCGZ_04588	scaffold160:101030-101541(+)	798	798	0	28	265	35	3	RcLhcb1.1 RcLhcb1.2 RcLhcb1.3	MeLhcb1.1 MeLhcb1.2 MeLhcb1.3	AtLhcb1.1 AtLhcb1.2 AtLhcb1.3 AtLhcb1.4 AtLhcb1.5	Lhcb1
	JcLhcb1.2	JCGZ_04587	scaffold160:97668-99902(-)	798	798	0	17	265	35	3	RcLhcb1.1 RcLhcb1.2 RcLhcb1.3	MeLhcb1.1 MeLhcb1.2 MeLhcb1.3	AtLhcb1.1 AtLhcb1.2 AtLhcb1.3 AtLhcb1.4 AtLhcb1.5	Lhcb1
	JcLhcb2	JCGZ_18481	scaffold529:239615-242657(-)	798	2225	1	10	265	37	3	RcLhcb2	MeLhcb2.1 MeLhcb2.2	AtLhcb2.1 AtLhcb2.2 AtLhcb2.3	Lhcb2
	JcLhcb3	JCGZ_00703	scaffold108:360030-361580(-)	804	1202	2	5	267	44	3	RcLhcb3	MeLhcb3	AtLhcb3	Lhcb3
	JcLhcb4	JCGZ_25025	scaffold843:295347-296708(+)	858	957	1	8	285	31	3	RcLhcb4	MeLhcb4	AtLhcb4.1 AtLhcb4.2	Lhcb4
	JcLhcb8	JCGZ_01281	scaffold11:738853-741122(+)	840	1786	1	0	279	31	3	RcLhcb8	MeLhcb8	AtLhcb8	Lhcb8
	JcLhcb5	JCGZ_06701	scaffold200:160436-162429(-)	876	1573	5	1	291	41	3	RcLhcb5	MeLhcb5	AtLhcb5	Lhcb5
	JcLhcb6	JCGZ_20203	scaffold645:375633-376811(+)	765	844	1	1	254	49	3	RcLhcb6	MeLhcb6	AtLhcb6	Lhcb6
	JcLhcb7	JCGZ_08016	scaffold221:188978-192485(-)	1014	3050	5	0	337	52	3	RcLhcb7	MeLhcb7	AtLhcb7	Lhcb7
PsbS	JcPsbS	JCGZ_12094	scaffold339:857142-859344(-)	825	1726	3	2	274	62	4	RcPsbS	MePsbS	AtPsbS	PsbS
ELIP	JcELIP	JCGZ_02231	scaffold119:1376949-1378080(-)	585	808	2	6	194	86	3	RcELIP	MeELIP	AtELIP1 AtELIP2	ELIP
OHP	JcOHP1	JCGZ_09105	scaffold250:2426033-2426931(-)	357	552	2	2	118	48	1	RcOHP1	MeOHP1	AtOHP1	OHP1
	JcOHP2	JCGZ_08332	scaffold224:432321-435041(+)	552	2339	1	0	183	47	1	RcOHP2	MeOHP2.1 MeOHP2.2	AtOHP2	OHP2
SEP	JcSEP1	JCGZ_23193	scaffold7:432788-436707(-)	438	3498	3	3	145	71	2	RcSEP1	MeSEP1	AtSEP1	SEP1
	JcSEP2	JCGZ_09398	scaffold255:592063-593949(+)	582	1322	1	0	193	46	2	RcSEP2	MeSEP2	AtSEP2	SEP2
	JcSEP3	JCGZ_03488	scaffold137:779011-781250(+)	780	1919	2	4	259	100	2	RcSEP3	MeSEP3.1 MeSEP3.2	AtSEP3.1 AtSEP3.2	SEP3
	JcSEP6	JCGZ_26324	scaffold906:2486216-2487759(-)	759	932	2	0	252	88	2	RcSEP6	MeSEP6	-	SEP6
	JcSEP4	JCGZ_06634	scaffold20:991705-992601(+)	570	570	0	0	189	56	2	RcSEP4	MeSEP4	AtSEP4	SEP4
	JcSEP5	JCGZ_07816	scaffold211:5206555-5210867(-)	447	3909	4	0	148	75	1	RcSEP5	MeSEP5	AtSEP5	SEP5
Psb33	JcPsb33	JCGZ_17235	scaffold5:830004-833169(+)	867	2803	2	1	288	62	1	RcPsb33	MePsb33.1 MePsb33.2	AtPsb33	Psb33
FCII	JcFCII	JCGZ_24260	scaffold813:61000-71385(-)	1500	9600	9	0	499	86	1	RcFCII	MeFCII	AtFCII	FCII

Table 2 Detailed information of 25 motifs identified in this study.

Motifs were identified using MEME.

Motif	E-value	Sites	Width	Best match
Motif 1	2.5e-1988	205	21	EJINGRLAMLGFLGFLVQEIL
Motif 2	6.2e-1170	24	60	ARNGVKFGEAVWFKAGAQIFSEGGLDYLGNPSLIHAQSILAIWACQVVLMGAVEGYRVAG
Motif 3	9.4e-992	69	23	YLDGELPGDYGFDPAGLSADPET
Motif 4	8.6e-894	64	21	TGKGPJENLADHLADPVHNNI
Motif 5	2.6e-582	58	21	GSFDPLGLADDPEAFAELKVK
Motif 6	3.0e-546	40	29	TLFVIELJLIGYVEFRRWADLDNPGSVYP
Motif 7	1.9e-374	32	21	SPWYGPDRVKYLGPFSGETPS
Motif 8	4.2e-331	8	73	KFVDPRWIGGTWDLKQFZKDGKTDWDAVIDAEAKRRKWLEENPESSSNDEPVVFDTSIIPWWAWIKRYHLPEA
Motif 9	1.4E-216	9	41	QNLAKNVAGDIIGTRTEAADVKSTPFQPYSEVFGLQRFREC
Motif 10	3.5E-205	12	48	KTLLFVAVAGVLLIRKNEDIETLKKLLDETTLYDKQWQATWKDZNPSS
Motif 11	9E-200	5	80	IENRSPAEGAYSEGLJNAKLTQDGCIVCPTTDSTFDLRTGAIKDWYPKNPVLRVLTPALRTLYVYPVKTDEENIYISLRG
Motif 12	1E-170	4	80	KVYVGMRYWHPFTEEAIEQIKRDGITKLVVLPLYPQFSISTSGSSLRLLESIFREDEYLVNMQHTVIPSWYQREGYIKAM
Motif 13	8.1E-166	13	29	DVGYPGGLWFDPLGWGSGSPEKVKELRTK
Motif 14	3.6E-159	27	15	SWYDAGKVEYFAGSS
Motif 15	1.8E-154	25	21	TVCVKADPDRPLWFPGSTPPE
Motif 16	5.4E-153	24	11	WAYATNFVPGK
Motif 17	7.6E-149	20	21	PSAPEVMGNGRVTMRKTVKKA
Motif 18	8.6E-142	12	41	TGKGLLAQLNJETGJPIYELEPLVLFNVLFALFAAINASKD
Motif 19	3.6E-140	80	11	MATSTLAASSS
Motif 20	6.1E-140	4	80	GAFHFIEPVWWRVGYSKLKGDTLDYLGIPGLHLAGSQGVIVIAICQALLMVGPEYARYCGIEALEPLGIYLPGDINYPGG
Motif 21	3.4E-142	8	57	GPEPLLGVZPFLINLLADPVIERLPYVGAFLVKPLEAFISLVPLPKVEEGLASYGGG
Motif 22	2E-98	5	53	PPPKPAAQVALDDKNVITLEFQRQKAKELQEYFKQKKLEETDQGPFFGFLGKN
Motif 23	4.3E-98	23	11	PLGEVTDPJYP
Motif 24	4.6E-92	4	57	AYQSRVGPVEWLKPYTDETIIELGRKGVKNLLAVPISFVSEHIETLEEIDVEYKELA
Motif 25	1.4E-106	10	29	NWVPAVPLAALPGGZATYJGQPVPTGLLP